Chapter 2: Sample Design & Fielding Procedures

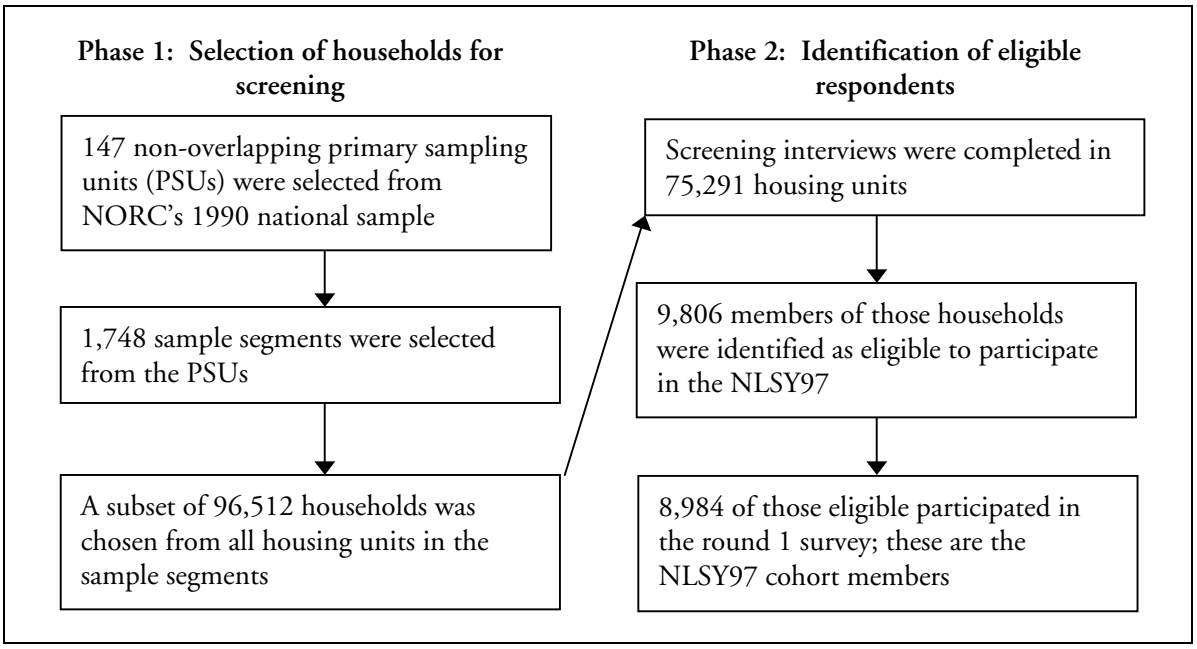
2.1 Sample Design & Screening Process

Sampling Procedures

The NLSY97 cohort comprises two independent probability samples: a cross-sectional sample and an oversample of black and/or Hispanic respondents. The cohort was selected using these two samples to meet the survey design requirement of providing sufficient numbers of black and Hispanic respondents for statistical analysis. Information on the actual size and racial/ethnic composition of the cohort is presented in section 1.3, "NLSY97 Sample."

The NLSY97 cohort was selected in two phases, as pictured in Figure 1. In the first phase, a list of housing units for the cross-sectional sample and the oversample was derived from two independently selected, stratified multistage area probability samples. This ensured an accurate representation of different sections of the population defined by race, income, region, and other factors. In the second phase, subsamples of the eligible persons identified in the first phase were selected for interview.

2.1 Figure 1. Selection of NLSY97 Respondents



The listing of eligible housing units was composed of 96,512 households, defined as a single room or group of rooms intended as separate living quarters for a family, for a group of unrelated persons living together, or for a person living alone. The list of housing units for each sample was selected in the following manner. First, 100 primary sampling units (PSUs)[1] for each sample were chosen from NORC's

[1] There are 100 PSUs in the cross-sectional sample and 100 PSUs in the oversample; however, some PSUs were selected in both samples. Thus, there are a total of 147 non-overlapping PSUs included in the NLSY97.

1990 national sample. In the cross-sectional sample, each PSU represented either a metropolitan area or one or more non-metropolitan counties with a minimum of 2,000 housing units. The supplemental sample defined PSUs differently from the cross-sectional sample; counties containing large percentages of minorities were merged to create areas containing a minimum of 2000 housing units. Second, regardless of sample, segments containing one or more adjoining blocks—and at least 75 housing units—were selected from each PSU. Finally, a subset of housing units within the segment comprised the NORC listing of households eligible for interview.

The second phase identified all NLSY97-eligible individuals age 12 to 16 as of December 31, 1996, in each household. NORC interviewers went to the households and administered a short interview called the simple screener, a portion of the *Screener, Household Roster, and Nonresident Roster Questionnaire,* which collected the age or date of birth of every person linked to a particular household. The survey collected these data for over 150,000 people. In cross-sectional sampling units, if the household included one or more occupants in the eligible age range, NORC interviewers asked those individuals to participate in the first NLSY97 interview. In supplemental sampling units, the interviewer continued with the extended screener, which established the race and ethnicity of household members. If a person of the correct age and of black or Hispanic race/ethnicity resided in the household, her or she was asked to participate in the survey. Any person in the above age range who completed the first round interview is considered a member of the NLSY97 cohort. Base year interviews were conducted between January and early October 1997 and between March and May 1998 (see section 2.2 for details). Of the 9,806 individuals selected for interview during household screenings, a total of 8,984 (91.6 percent) were interviewed.

During the NLSY97 screening process, two additional nationally representative samples were identified to participate in the administration of the *CAT-ASVAB*. The first group, the Student Testing Program (STP), consisted of students who expected to be in the 10^{th} through 12^{th} grades in the fall of 1997. Included were many respondents who also participated in the main NLSY97 survey, as well as youths who refused to participate in or were not eligible for the NLSY97. The second sample, the Enlistment Testing Program (ETP), was a nationally representative sample of youths 18 to 23 years old as of June 1, 1997. This group provided the normative information that will be used by the Department of Defense to determine the score distribution of military-eligible youths and will help to assess the impact of these tests on minority and female military eligibility.

Cross-Sectional Sample

For the cross-sectional sample, 54,179 screening interviews were carried out among 1,149 sample segments in 100 primary sampling units (PSUs), drawn from the NORC master probability sample of the United States. The cross-sectional screening established three samples:

(1) **Main NLSY97 Sample:** A cross-sectional sample designed to be representative of young people living in the United States during round 1 and born January 1, 1980, through December 31, 1984. This sample is designed to maximize the statistical efficiency of samples through the several stages of sample selection (counties, enumeration districts, blocks, sample listing units). Probabilities of selection are based upon total housing units in a geographic area.

Following the initial screening process, 7,327 individuals from the cross-sectional sample were designated to be interviewed in the NLSY97 survey; of those, 92.1 percent, or 6,748 respondents, completed the round 1 interview.

(2) **Department of Defense Student Testing Program (STP) Sample:** A nationally representative sample of students living in the United States during round 1 and born June 2, 1973, through December 31, 1984, who—depending on the time of the household screening—were in grades 9–11 in the spring or summer of 1997, were not enrolled during the spring and summer but expected to be in grades 10–12 in the fall of 1997, or were enrolled in grades 10–12 during the fall of 1997. (See the "Administration of the *CAT-ASVAB*" section of this guide for more information.) Some NLSY97 respondents were also eligible for the STP sample.

(3) **Department of Defense Enlistment Testing Program (ETP) Sample:** A cross-sectional sample designed to be representative of the noninstitutionalized segment of young people living in the United States during round 1 and born June 2, 1973, through June 1, 1979.

Supplemental Sample

Statistically efficient samples of black and Hispanic respondents were created by oversampling these minorities in 100 PSUs in NORC's national sample. For the supplemental sample, 21,112 screening interviews were conducted in 599 sample segments. The supplemental screening produced three samples:

(1) **NLSY97 Black and Hispanic Oversample:** A supplemental sample designed to oversample Hispanic and black respondents living in the United States during round 1 and born January 1, 1980, through December 31, 1984. Stratification specifically relevant for Hispanics and blacks was used. Oversample respondents were chosen with a probability based on size measures for these groups rather than for the general population. This should make it possible to equalize the distribution of the targeted groups among the various sampling units more than would otherwise be the case.

After screening, 2,479 individuals from the supplemental sample were designated for interview in the NLSY97, and of these, 90.2 percent or 2,236 respondents completed the round 1 interview.

(2) **Department of Defense STP Sample:** A nationally representative sample of students, selected regardless of race and/or ethnicity, living in the United States during round 1 and born June 2, 1973, through December 31, 1984. Members of this sample are those who—depending on the time of the household screening—were in grades 9–11 in the spring or summer of 1997, were not enrolled during the spring and summer but expected to be in grades 10–12 in the fall of 1997, or were enrolled in grades 10–12 during the fall of 1997.

(3) **Department of Defense ETP Black and Hispanic Oversample:** A sample of black or Hispanic youths living in the United States during round 1 and born June 2, 1973, through June 1, 1979.

Data hint ➔ Users can identify the cross-sectional or supplemental sample type of each respondent by referring to the sample type variable (CV_SAMPLE_TYPE—R12358.) on the NLSY97 CD-ROM.

Screening Procedures

The screening interview was completed by NORC in 75,291 housing units. These interviews occurred in 1,748 sample segments of 147 non-overlapping PSUs, including most of the fifty states and the District of Columbia.[2] The screening interview was designed to elicit information allowing identification of household occupants eligible for inclusion in the NLSY97 sample. The NLSY97 screening interviews were completed within 94.1 percent of the cross-sectional and 93.1 percent of the supplemental occupied housing units selected for screening. Table 1 presents a summary of completed interviews in round 1.

2.1 Table 1. NLSY97 Round 1 Interview Completion

Sample	Eligible for interviewing	Interviewed round 1	
Total Cohort	9806	8984	91.6%
Cross-Sectional Sample	7327	6748	92.1%
Supplemental Sample	2479	2236	90.2%

Sampling procedures were developed to establish links between housing units in the sample PSUs and individuals who might be temporarily absent. As part of the screening process, household informants were asked if there were any persons for whom the housing unit was the usual place of residence, but who were away from the housing unit at the time of the survey. Included in this group were college students, persons in the military, and persons in prisons or other institutions. Sampling procedures were also established for those residing in a selected housing unit whose usual place of residence was elsewhere. Table 2 lists the NLSY97 status (e.g., included in the sample, excluded, or restricted) for youths not in their usual place of residence at the time of the survey.

[2] There are 100 PSUs in the cross-sectional sample and 100 PSUs in the oversample; however, some PSUs were selected in both samples. Thus, there are a total of 147 non-overlapping PSUs included in the NLSY97.

2.1 Table 2. NLSY97 Sampling Status of Youths by Housing Arrangement

Housing arrangement	Status
Exchange students	Included if the youth lived in the sample housing unit for at least six months during 1997.
Youths whose temporary residence was a group quarters structure (e.g., prisons, boarding school, college dormitories)	Included if their usual place of residence was in a selected PSU. Excluded otherwise.
Youths whose usual place of residence was not in a selected PSU, but whose temporary residence was within a PSU	Excluded.
Youths in a foreign school	Included.
Youths linked to two or more housing units	If the respondent's mother is alive and her housing unit is in a sample housing unit, the youth is linked there. Otherwise, the youth is linked to the father's housing unit. If neither the mother nor the father is alive and living in a sample housing unit, the youth is linked to one of the sample housing units at random.
Youths who cannot be linked to any other housing unit	Included if the youth is residing at a sample housing unit when the screening interview is conducted.

Siblings: The NLS sample design, which selected every eligible person connected to the housing unit, generated a sample of siblings living in the same housing unit and satisfying the NLSY97 age restrictions. However, the NLSY97 samples do not contain nationally representative samples of siblings of all ages and living arrangements. Care should be used in generalizing from the findings of sibling studies based on the NLSY97. See Table 3 in section 1.3 for the numbers of sibling groups in the NLSY97.

Other technical information on the sample assignment process can be found in (1) the *Field Interviewer Reference Manual,* which includes a copy of the screening instrument, and (2) the *Technical Sampling Report*, which describes the NLSY97 sample selection procedures for both subsamples. Contact NLS User Services concerning the availability of these documents.

2.2 Interview Methods

This section first discusses the data collection methods used for the five round 1 survey instruments: the *Screener, Household Roster, and Nonresident Roster Questionnaire;* the *Youth Questionnaire;* the *Parent Questionnaire;* the *School Survey*; and the *CAT-ASVAB*. Following this overview, the section briefly describes interview administration in subsequent survey rounds. The content of these instruments is described in section 1.4, "Content of the NLSY97."

Users should note that respondents have received $10 for their participation in rounds 1–3, and responding parents received $10 when they completed the round 1 interview. In round 4, survey administrators offered different levels of incentives to respondents in an effort to study the effects of incentive level on survey participation. Three levels of compensation were offered: $10, $15, and $20. In

addition, half of the respondents at each level were paid in advance and half were paid upon completion of the interview. Both the level and the timing of the compensation are included in the variable PAYINCENT, found on the round 4 CD–ROM.

The field periods have differed somewhat across rounds. Table 1 indicates when the first several rounds were fielded, along with the total response rate.

2.2 Table 1. NLSY97 Sample Sizes, Retention Rates, and Fielding Periods

Round	Fielding period	Cross-sectional sample		Supplemental sample		Total sample	
		Total	Retention rate	Total	Retention rate	Total	Retention rate
1	February–October 1997 and March–May 1998	6748	—	2236	—	8984	—
2	October 1998–April 1999	6279	93.0	2107	94.2	8386	93.3
3	October 1999–April 2000	6173	91.5	2036	91.1	8209	91.4
4	November 2000–May 2001	6055	89.7	2026	90.6	8081	89.9
5	November 2001–May 2002[1]	NA	NA	NA	NA	NA	NA

Note: Retention rate is defined as the percentage of base year respondents remaining eligible who were interviewed in a given survey year; deceased respondents are included in the calculations.
[1] Round 5 was fielded in 2001–02 but is not discussed in this guide.

Round 1 Interview Methods

Fielding Period: Most round 1 NLSY97 interviews were conducted between January and early October 1997. Due to concerns about the number of eligible youths found during the initial field period, investigators decided to conduct a refielding between March and May 1998. During this second part of the initial survey round, 395 additional respondents were interviewed. These respondents were administered the same instrument as those initially interviewed in 1997. See section 2.3 for more information about the composition of the NLSY97 sample.

Data hint ➔ Respondents selected for the NLSY97 sample during the refielding are identified by the refielding symbol (CV_REFIELD_YOUTH).

Researchers analyzing topics where time periods are critical should carefully examine the reference period of the questions, as well as the actual interview date for individual respondents. In particular, the round 1 fielding period has implications for questions on education; see section 4.2.2, "Educational Status & Attainment," for more information.

Researchers should also pay close attention to the elapsed time between interviews for each respondent. While the time between the first and second interviews was about 18 months for most respondents, it may be somewhat less for those first interviewed during the refielding period.

Data hint ➔ The respondent's interview date for each round can be identified by using three created variables: CV_INTERVIEW_DATE_D, CV_INTERVIEW_DATE_M, and CV_INTERVIEW_DATE_Y.

Screener, Household Roster, and Nonresident Roster Questionnaire

Choice of household informant: To identify youths potentially eligible for the NLSY97, the screener collected data from selected households within a sample area. A single member of the household, designated as the household informant, was asked to provide certain information on persons who usually resided in the household. To ensure more accurate reporting of these data, the NLSY97 required the household informant to be age 18 or older and to consider the selected household his or her usual place of residence.

Computer-Assisted Personal Interview (CAPI): After a household informant was chosen to complete the *Screener, Household Roster, and Nonresident Roster Questionnaire,* interviewers used a CAPI system to collect data. Computer software automatically guided interviewers through an electronic questionnaire, selecting the next question based on a respondent's answers. The program also prevented interviewers from entering invalid values and warned interviewers about implausible answers. A set of checks within the CAPI system lowered the probability of inconsistent data both during an interview and over time. To ensure that accurate data were collected from Spanish-speaking respondents, CHRR prepared both English and Spanish versions of all survey instruments, and NORC employed bilingual Spanish-speaking interviewers to administer the Spanish version to those requesting it. During the initial round, the Spanish version of the questionnaire was requested by 297 responding parents and 96 NLSY97 youths.

Screen and Go: In round 1, use of the computer-assisted personal interviewing system (CAPI) allowed for a screen and go method of screening households. When an NLSY97-eligible youth was identified in the simple screener portion of the interview, information from the remainder of the *Screener, Household Roster, and Nonresident Roster Questionnaire* was collected. Selected data (e.g., basic demographic information, a roster of household members) were then transferred automatically into the *Parent* and *Youth Questionnaires* for verification and use during the interview. Therefore, the interviewer could administer the parent or the youth portion of the NLSY97 immediately. It was expected that this would increase the likelihood that eligible youths participated in the survey since the number of visits interviewers had to make to a household decreased.

However, in some cases, the respondents (parent and youth) were not available to participate in the parent and youth interviews immediately after screening. In these cases, a screen and come back method was utilized, in which the interviewer made an appointment to return to the household to administer the *Youth* and *Parent Questionnaires* at a convenient time.

Paper Screener: During round 1, the interviewers had the option of using a paper screener to perform the initial screening of the household. The paper screener collected the same basic information as the initial CAPI screener. This was useful in cases where the simple screener information could not be collected using CAPI (e.g., weather conditions, computer battery life, dangerous neighborhood) and also gave the interviewer an alternative medium for collecting the initial screener data. Like the screen and go model, the paper screener was designed to determine if anyone residing in the housing unit was eligible for either the NLSY97 or the administration of the *CAT-ASVAB*. If a youth was identified as being potentially eligible for the NLSY97, the information from the paper screener was entered into CAPI. The interviewer could then continue in CAPI with the *Screener, Household Roster, and Nonresident Roster Questionnaire* and the *Youth* and *Parent Questionnaires*. Approximately 28,000 paper screeners were administered, including those used for the screen and come back method described above.

Proxy Screener: In cases where a round 1 interviewer made several visits to a household and still had difficulty contacting household members to administer the initial screener, a proxy screener was administered to an adult living either next door to or directly across from the selected housing unit. Before the interviewer could administer a proxy screener, at least three attempts were made by the interviewer, on different days and at different times, to contact anyone in the selected housing unit.

The purpose of the proxy screener, a paper questionnaire, was to assess whether a person eligible for the NLSY97 resided in the household. In particular, the proxy screener was designed to determine the best time to establish contact with a household member, whether or not a person between the ages of 8 and 28 currently lived in the household, and the steps required to contact a household member. The broad 8–28 age range was intended to ensure that youths close to the endpoints of the actual age range were not missed due to inaccurate reporting. If the proxy screener indicated that none of the household members were in the age range of 8 to 28, the screener was coded as a proxy screener and no more attempts were made to contact the household. However, if the proxy informant was unable to definitively deny the presence of residents ages 8–28, the interviewer was instructed to return as many times as reasonable and necessary to administer the simple screener and, if appropriate, the remainder of the survey instruments. A total of 5,175 proxy screeners determined that no one between ages 8 and 28 lived in the household.

Gatekeepers: The gatekeeper disposition code was used in cases where the interviewer could not gain direct access to the sample household, such as a high-rise building with a locked door where access was

denied by a building manager or a gated housing community where the entry guard refused entrance. In these cases, the interviewer asked the gatekeeper or other community official whether anyone between the ages of 8 and 28 lived in the sample households. If the gatekeeper was unable to definitively deny the presence of household members ages 8–28, the interviewer then attempted to gain access to the household in order to complete the *Screener, Household Roster, and Nonresident Roster Questionnaire* and was not permitted to use this disposition code. A total of 4,055 cases were closed with a gatekeeper disposition code after the interviewer determined that no one between ages 8 and 28 lived in the household. This code was mainly used in gated housing communities for senior citizens.

Telephone Screener: In rare cases at the conclusion of the field period, the simple screener was conducted by telephone. A total of 931 telephone screeners were administered. Instances in which the housing unit was contacted by telephone include:

(1) The proxy screener revealed a person between the ages of 8 and 28 living in the household and the interviewer was unable to contact anyone in the housing unit on three subsequent in-person visits; or

(2) The interviewer made three in-person visits but was unable to find a neighbor to whom he or she could administer the proxy screener.

The full *Screener, Household Roster, and Nonresident Roster Questionnaire* was also administered by telephone in rare instances. Situations in which the full instrument was conducted by telephone include:

(1) After completing the paper screener, the interviewer was unable to contact anyone in the housing unit to complete the full extended screener. At least three in-person contacts must have been attempted before the telephone contact was approved.

(2) The sample housing unit was inside a residential community to which the interviewer was barred access by the community (e.g., housing board authority). Prior to the telephone interview, the correct person must have been contacted about gaining access at least three times (in person, by telephone, or by letter).

NLSY97 *Parent Questionnaire* and *Youth Questionnaire*

When the *Screener, Household Roster, and Nonresident Roster Questionnaire* was complete, any NLSY97-eligible youth(s) and one of the youth's parents (the responding parent) were interviewed using CAPI. Prior to these interviews, selected data (e.g., basic demographic information, a roster of household members) were automatically transferred into the *Parent Questionnaire* and the *Youth Questionnaire* for verification and use during the interviews. Consequently, the interviewer was able to administer the parent or the youth portion of the NLSY97 immediately. CAPI interviews were conducted in either English or Spanish; parent and youth respondents could choose either version.

Data hint ➔

In round 1, the NLSY97 youth respondent(s) and responding parent(s) in the household are listed on the household roster, but they are referred to as "Household Member #" in the same way as noninterviewed household members. The youth respondent's position on the household roster can be identified by using the variable YOUTH_HHID.01. The responding parent's position on the roster is provided in PARYOUTH_PARENTID. See section 4.6.5, "Household Composition," for further discussion of the structure and use of the household roster.

Choice of Parent: One parent of each respondent was asked to participate in the parent interview. This parent was identified during the household roster portion of the survey. The responding parent (or guardian) was asked for extensive background information, including marital and employment histories. He or she was also asked to answer questions about the family in general, as well as to provide information about aspects of his or her (NLSY97-eligible) children's lives.

The choice of the preferred responding parent was based on the pre-ordered list in Figure 1. For example, a biological mother was chosen before a biological father, and so forth. However, in some cases a parent figure lower on the list was chosen if a parent higher on the list was in the household but was not available at the time of the interview. If the youth did not live with a parent-type figure, or lived with a guardian or parent not listed, no parent was interviewed; the youth's record will not contain any data from the *Parent Questionnaire*. Users should note that the records of some youths who do live with a listed parent or parent-figure do not contain any data from the *Parent Questionnaire* due to nonresponse.

2.2 Figure 1. Priority for Choosing Responding Parent

1	Biological mother
2	Biological father
3	Adoptive mother
4	Adoptive father
5	Stepmother
6	Stepfather
7	Guardian, relative
8	Foster parent, youth lived with for 2 or more years
9	Other non-relative, youth lived with for 2 or more years
10	Mother-figure, relative
11	Father-figure, relative
12	Mother-figure, non-relative youth lived with for 2 or more years
13	Father-figure, non-relative youth lived with for 2 or more years

Interviews are available with 6,124 parents; 7,942 youth respondents have information available from a parent interview. Table 2 shows the number of respondents by age who had a parent participate in the round 1 survey.

2.2 Table 2. NLSY97 Youths by Age and Parent Interview Availability

Age (birth year)	Total number of youths	Youths with a parent interview
12 (1984)	1771	1583 (89.4%)
13 (1983)	1807	1615 (89.4%)
14 (1982)	1841	1595 (86.6%)
15 (1981)	1874	1668 (89.0%)
16 (1980)	1691	1481 (87.6%)
Total	**8984**	**7942 (88.4%)**

Note: Table based on R05367. and R07359.

In multiple respondent households, more than one parent may have been interviewed during round 1 if the selection criteria above indicated different parents for different NLSY97-eligible youths in the household. For example, if a couple residing in a sample household each had an NLSY97-eligible youth from a previous marriage, the biological parent of each youth would be interviewed. The survey first collected parent-specific information from each parent and then asked for information about the NLSY97-eligible youth matched to that parent. In this example, each parent would be asked to provide youth-specific information only for his or her NLSY97-eligible biological child.

Due to a computer programming error, however, both parents in some multiple respondent households were asked to provide youth-specific information only for the oldest NLSY97-eligible youth(s) living in the household. In the example above, both parents would be asked to give information about the older of the two children. In these infrequent instances, the correct parent-specific information is matched to each youth, but one or more youths in the household do not have any youth-specific information. This programming error was corrected during the survey period and affected only 33 youth cases.

Audio Computer-Assisted Self-Interview (ACASI): The parent and youth portions of the NLSY97 survey used an audio computer-assisted self-interview (ACASI) to obtain potentially sensitive information. The respondent was able to listen to the questions with earphones or turn off the audio and read the questionnaire from the computer screen. Compared to traditional paper-and-pencil self-administered sections, the computerized version permits more complex questionnaire structuring, and the audio component theoretically improves response quality when the respondent's literacy is in question. As with the interviewer-administered instruments, the ACASI was available in Spanish or English.

User Notes: Each NLSY97 questionnaire includes an interviewer remarks section, which interviewers complete after finishing the interview with the respondent. This section records objective information about the interview, such as the presence of another person during the survey, where the interview took place, and the language in which the questionnaire was administered. Interviewers are also asked to provide an assessment of the interview, stating how cooperative the youth was, how well the youth appeared to understand the questions, whether the youth seemed to be candid and honest, and whether there were any special circumstances that might affect the quality of the data (e.g., respondent lacks social skills, has a mental impairment, has a physical disability, is under the influence of alcohol or drugs). Finally, the interviewer observes the youth's home and neighborhood environment, describing the interior and exterior condition of the youth's home, the type of neighborhood (rural and agricultural, suburban residential, urban residential, urban mixed residential and commercial, etc.), the type of residence most common on the youth's street, and whether the interviewer felt safe in the youth's neighborhood. These questions help survey staff to plan for future interviews by anticipating potential problems and provide researchers with a general idea of the quality of the respondent's answers. Questions found in the interviewer remarks section have the prefix "YIR" in their question name.

Supplemental NLSY97 Studies

School Survey (1996). Designed with an emphasis on the school-to-work transition, round 1 of the NLSY97 also included a mail survey of schools. Principals (or their proxies) were asked to complete a self-administered instrument that focused on institutional-level attributes such as school policies and management as well as student-level "experience" data. See section 4.2.5, "School & Transcript Surveys," for more detail about the content of the survey.

Schools in the NLSY97 sample areas that had a 12th grade comprised the sample for this survey. As depicted in Figure 1 in section 2.1 of this chapter, the NLSY97 sample was drawn from 147 primary sampling units (PSUs).[3] The PSUs were further divided into sample segments. All schools in any county with a segment selected for NLSY97 sampling were included in the survey. There were some counties in the PSUs from which no sample segments were selected. The 1996 survey did not include schools in these counties. Schools were identified using the Quality Education Data (QED) file, a proprietary national database of primary and secondary schools in the United States.

[3] There are 100 PSUs in the cross-sectional sample and 100 PSUs in the oversample; however, some PSUs were selected in both samples. Thus, a total of 147 non-overlapping PSUs are included in the NLSY97.

The original school survey form was mailed in September 1996; in-scope schools that did not respond by December 1996 were sent a shorter version of the survey, the "critical items" questionnaire. Of the 7,390 in-scope schools that received the survey, 5,295 responded to either the original school survey or the critical items questionnaire. The response rate by the end of the field period, April 5, 1997, was 71.6 percent.

Answer forms for the original school survey were electronically scanned by NORC. However, some hand editing was necessary. The majority of the edited questions were in decimal format. To ensure clean data, the answers were verified by randomly selecting cases, keying the data, and comparing the keyed data files against the scanned data files. The critical items questionnaire did not use a scannable format; the data were keyed using Computer Assisted Data Entry (CADE) and verified twice.

CAT-ASVAB: From summer 1997 through spring 1998, most NLSY97 respondents were administered the computer adaptive version of the *Armed Services Vocational Aptitude Battery (CAT-ASVAB),* as well as the *Interest-Finder*. See section 4.1.2, "Administration of the *CAT-ASVAB*," for more information.

Rounds 2–4 Interview Methods

Fielding Periods: The round 2 survey was conducted from October 1998 through April 1999. Most respondents were surveyed approximately 18 months after their first interview, although the elapsed time between interviews is substantially less for some respondents. The round 3 survey was conducted from October 1999 through April 2000. Round 4 surveys were administered from November 2000 through May 2001.

Locating respondents is a coordinated effort of NORC's central office, locating shop, and local-level field staff. Prior to fielding, NORC's central office sends a short, informative "locator letter" to each respondent reminding him or her of the upcoming interview and confirming the respondent's current address and phone number.

Youth Questionnaire: As in round 1, the interviews were conducted using a CAPI instrument, administered in person by an interviewer with a laptop computer. During sensitive portions of the interview, the respondents entered their answers directly into the laptop rather than interacting with the interviewer. This self-administered portion, called ACASI, included an audio option so that the respondents could listen to the questions and answers being read via headphones if they preferred.

Household Income Update: This brief questionnaire collected basic income information from one of the respondent's parents (usually the parent who signed the youth's interview consent form). All respondents who live with a parent are eligible for this questionnaire, regardless of age or other criteria for

independence. The parent answered these questions on a self-administered paper instrument. Interviewers then entered the data into a computer-assisted questionnaire on their laptops and attached the information to the records of all NLSY97 youths in the household. Additional quality control checks were performed in the central office, where hard copy questionnaires were reviewed against the coded data. In round 2, parents of 7,601 respondents answered at least one question from the *Household Income Update*; parents of 5,488 respondents answered at least one question in round 3; and 5,225 parents of respondents answered at least one question in round 4.

Transcript Survey. In winter 1999–2000, the 2000 NLSY97 transcript survey sought high school transcripts for all sample respondents who were no longer enrolled in high school and for whom field interviewers had secured parent and respondent consent for transcript release. Eligible respondents were those who either had graduated from high school or who were age 18 or older and no longer enrolled in high school. Transcripts were received and processed for 1,417 respondents. Using course catalogs, transcript data, and clarification calls to school administrators, survey staff constructed histories of courses taken and term enrollment calendars for each youth. Data files also include information on absences, standardized test scores, and indicators of special education, gifted/talented, and high school graduation status. Courses were coded into the Revised Secondary School Taxonomy (SST-R). Public use data will be available on the round 4 Event History CD-ROM.

School Survey (2000). Round 3 of the NLSY97 also included a repeat survey of schools. Principals (or their proxies) were asked to complete a self-administered instrument similar to that used in 1996. To reduce the time burden, questionnaire items from the 1996 instrument were modified to encourage respondents to provide approximate values rather than requiring them to consult administrative records for exact figures. See section 4.2.5, "School & Transcript Surveys," for more detail about the content of the survey.

As in 1996, schools in the NLSY97 PSUs that had a 12th grade were mailed survey instruments. However, the 2000 sample was expanded to include vocational schools. The sample also included schools in the counties that were in NLSY97 PSUs but did not include any sample segments. Schools in these counties had been omitted from the 1996 survey but were included for limited data collection in 2000. No telephone follow-up was done for schools in these "omitted counties." Finally, in addition to the geographically based sample, other schools were included if an NLSY97 respondent was enrolled during round 2 and that school met the grade and program requirements for eligibility. Schools were identified using the 1998 Quality Education Data (QED) file.

By January 2000, survey staff had secured cooperation from state school officers and local school districts. In February 2000, questionnaires were mailed to 9,632 sampled schools, including 8,925 schools in a

longitudinal sample (comparable to the 1996 school survey), 492 in the omitted counties sample, and 215 eligible only due to round 2 youth enrollment. After mail and telephone follow-up, 5,955 schools (71.6 percent) in the longitudinal sample (comparable to the 1996 school survey) completed questionnaires. The overall response rate for all schools in the 2000 survey was 71 percent.

Due to "births" and "deaths" of schools between 1996 and 2000 and nonresponse in 1996, not all schools in the longitudinal sample are present in the 1996 data. The retention rate of 1996 schools into the 2000 survey was 74.2 percent (3,900 of 5,253).

Validation Reinterviews. After each round of the NLSY97, validation reinterviews are conducted with randomly selected respondents in order to confirm that their interviews were administered as reported by the interviewer and to solicit feedback on interviewers' conduct. Most validations are conducted over the telephone by the NORC phone center, with a small number conducted in person or by mail. These data offer opportunities for studying response variance, item reliability, and other methodological issues. Though these reinterviews have been administered each year since round 2, only the round 4 data have been released for public users to date. These variables have "VALIDR4" as the beginning of each question name and are found on the round 4 main file CD-ROM.

Between November 2000 and July 2001, 989 respondents completed validation reinterviews for round 4. This produced an overall project validation rate of 12.2% of completed interviews. The short telephone questionnaire included a validation component that asked for details about the respondents' original round 4 interview (e.g., duration, mode) and information on whether or not they were paid for their participation. The reinterview component involved re-asking questions that were drawn directly from the youth interview. This component included some characteristics of their current residence, several expectations questions, a question about weekly family activities, and two questions concerning the respondent's income from the previous year. Comparable to similar questions from the main interview data, these re-interview data are chosen to represent a variety of question types with different response variance characteristics. Finally, respondents are asked whether the interviewer they had in round 4 was the same one who conducted their interview in round 3.

2.3 Sample Size & Composition

For more information about the representativeness of the sample members, users should consult the *NLSY97 Technical Sampling Report* (2000). Although fewer age-eligible youths than expected were found during the household screenings, no correlation has been identified between education, income, area of residence, etc., and participation in the survey.

Of the youths eligible for interview in the first round, 8,984 were actually interviewed. Table 1 illustrates the racial, ethnic, and gender composition of the initial sample and the respondents participating in subsequent rounds.

2.3 Table 1. Racial, Ethnic & Gender Composition of NLSY97 Sample

Gender	Race/Ethnicity				
	Black	Hispanic	Non-black/ non-Hispanic	Mixed	Total
Round 1					
Male	1169	977	2413	40	4599
Female	1166	924	2252	43	4385
Total	2335	1901	4665	83	8984
Round 2					
Male	1103	904	2238	38	4283
Female	1101	868	2095	39	4103
Total	2204	1772	4333	77	8386
Round 3					
Male	1062	876	2193	39	4170
Female	1071	853	2076	39	4039
Total	2133	1729	4269	78	8209
Round 4					
Male	1065	862	2153	37	4117
Female	1059	837	2027	41	3964
Total	2124	1699	4180	78	8081

Note: Table based on KEY!RACE_ETHNICITY (R14826.), KEY!SEX (R05363.), and RNI (R25102. and R38297.).

User Notes: The initial NLSY97 data release contained records for 9,022 respondents. However, an evaluation of the round 1 data revealed that 38 of these respondents either were not age-eligible for the cohort or were duplicates. The records of these out-of-scope respondents have been removed from the data, and numbers in this guide have been updated to reflect the new sample size of 8,984 respondents. Identification numbers of dropped respondents are included in the round 1 *NLSY97 Codebook Supplement* and are available from NLS User Services.

2.4 Retention and Reasons for Noninterview

After the initial survey round, some sample members do not respond to one or more subsequent interviews. Table 1 shows the retention rates by sample type for rounds 2, 3, and 4 of the NLSY97.

2.4 Table 1. Retention Rates by Sample Type and Gender

Sample	Round 2		Round 3		Round 4	
	# interviewed	Retention rate	# interviewed	Retention rate	# interviewed	Retention rate
Cross-sectional	**6279**	**93.0%**	**6173**	**91.5%**	**6055**	**89.7%**
Male	3213	92.9	3144	90.9	3098	89.6
Female	3066	93.2	3029	92.1	2957	89.9
Supplemental	**2107**	**94.2**	**2036**	**91.1**	**2026**	**90.6**
Male	1070	93.9	1026	90.0	1019	89.4
Female	1037	94.6	1010	92.2	1007	91.9
Total	**8386**	**93.3**	**8209**	**91.4**	**8081**	**89.9**

Note: Table based on RNI (R25102. and R38297.), KEY!SEX (R05363.), and CV_SAMPLE_TYPE (R12358.). Retention rate is defined as the percentage of all base-year respondents participating in a given survey. Deceased respondents are included in the calculations.

For each respondent who is not interviewed in a given round, NORC personnel assign a reason for noninterview code, contained in the variable RNI. Tables 2–4 summarize the reasons for noninterview among NLSY97 respondents during rounds 2, 3, and 4.

2.4 Table 2. Reason for Noninterview by Gender

Reason for noninterview	Deceased	Not locatable	Technical problem	R too ill	R unavailable	Refused interview	Other	Total
Round 2 total	7	104	6	6	42	428	5	**598**
Male	3	52	3	3	22	229	4	**316**
Female	4	52	3	3	20	199	1	**282**
Round 3 total	16	192	2	1	51	510	3	**775**
Male	7	107	2	1	34	275	3	**429**
Female	9	85	—	—	17	235	—	**346**
Round 4 total	15	172	6	6	80	612	12	**903**
Male	6	87	—	2	53	326	8	**482**
Female	9	85	6	4	27	286	4	**421**

Note: Table based on RNI (R25102. and R38297.) and KEY!SEX (R05363.).

2.4 Table 3. Reason for Noninterview by Sample Type

Reason for noninterview	Deceased	Not locatable	Technical problem	R too ill	R unavailable	Refused interview	Other	Total
Round 2 total	7	104	6	6	42	428	5	**598**
Cross-sectional	6	63	3	6	37	350	4	**469**
Supplemental	1	41	3	—	5	78	1	**129**
Round 3 total	16	192	2	1	51	510	3	**775**
Cross-sectional	13	121	2	1	35	400	3	**575**
Supplemental	3	71	—	—	16	110	—	**200**
Round 4 total	15	172	6	6	80	612	12	**903**
Cross-sectional	12	106	5	5	61	496	8	**693**
Supplemental	3	66	1	1	19	116	4	**210**

Note: Table based on RNI (R25102. and R38297.) and CV_SAMPLE_TYPE (R12358.).

2.4 Table 4. Reason for Noninterview by Race/Ethnicity

Reason for noninterview	Deceased	Not locatable	Technical problem	R too ill	R unavailable	Refused interview	Other	Total
Round 2 total	7	104	6	6	42	428	5	**598**
Non-black/non-Hisp.	2	22	2	3	22	278	3	**332**
Black	4	39	—	1	8	79	—	**131**
Hispanic	1	40	4	2	11	69	2	**129**
Mixed	—	3	—	—	1	2	—	**6**
Round 3 total	16	192	2	1	51	510	3	**775**
Non-black/non-Hisp.	8	65	1	1	23	297	1	**396**
Black	6	59	—	—	13	123	1	**202**
Hispanic	2	67	1	—	14	87	1	**172**
Mixed	—	1	—	—	1	3	—	**5**
Round 4 total	15	172	6	6	80	612	12	**903**
Non-black/non-Hisp.	6	61	1	5	33	375	4	**485**
Black	8	44	3	1	21	128	6	**211**
Hispanic	1	66	1	—	26	106	2	**202**
Mixed	—	1	1	—	—	3	—	**5**

Note: Table based on RNI (R25102. and R38297.) and KEY!RACE_ETHNICITY (R14826.).

2.5 Sample Weights & Design Effects

Sample Weights

The sampling weights, which are constructed in each survey year, provide the researcher with an estimate of how many individuals in the United States are represented by each respondent. Weighting decisions for the round 1 NLSY97 data were guided by the following principles. Individual case weights were assigned to produce group population estimates when used in tabulations. The assignment of individual respondent weights involved at least three types of adjustment. Interested users should consult the *NLSY97 Technical Sampling Report* for a step-by-step description of the following adjustment process.

Adjustment One: The first weighting adjustment involves the reciprocal of the probability of selection. Specifically, this probability of selection is a function of the probability of selection associated with the housing unit in which the respondent was located as well as the subsampling (if any) applied to individuals identified in screening.

Adjustment Two: This process adjusts for differential response (cooperation) rates in the screening phase. Differential cooperation rates are computed (and adjusted) based on geographic location, group membership, and within-group subclassification.

Adjustment Three: This weighting adjustment attempts to correct for certain types of random variation associated with sampling as well as sample "undercoverage." These ratio estimations are used to conform the sample to Census Bureau estimates of population totals.

Sampling Weights and Readjustments: NORC recalculates the sampling weights for all interviewed respondents after each survey round. These readjustments correct for differential nonresponse. The weights are created using base year sample parameters in a procedure similar to that described above. However, in the final stage of post-stratification, the weights are computed on the basis of completed cases in that survey year rather than on the number of respondents in the entire sample.

User Notes: Various sampling weights have been created in different survey years. The figure below shows the variables created in each round and the question name of each variable. Cross-sectional weights refer to the cross-sectional sample. Panel weights include only those respondents who have been interviewed in every round up to that round's interview date. The cumulative cases method refers to a new method for creating more accurate sampling weights.

2.5 Figure 1. Sampling Weight Variable for All Rounds

Sampling Weight Variables	Round 1	Round 2	Round 3	Round 4
Sampling Weight (includes round 4 "old method")	SAMPLING_WEIGHT	SAMPLING_WEIGHT	SAMPLING_WEIGHT	SAMPLING_WEIGHT
Cross-Sectional Sampling Weight	CS_SAMPLING_WEIGHT	CS_SAMPLING_WEIGHT	CS_SAMPLING_WEIGHT	--
Sampling Weight Cumulative Cases Method	--	--	--	SAMPLING_WEIGHT_CC
Sampling Weight Panel Method	--	--	SAMPLING_PANEL_WEIGHT_R3	SAMPLING_PANEL_WEIGHT
Cross-Sectional Panel Weight	--	--	CS_PANEL_WEIGHT	--

Practical Usage

Researchers should weight the observations using the weights provided if tabulating sample characteristics in order to describe the population represented (i.e., computing sample means, totals, or proportions). The use of weights may not be appropriate without other adjustments for the following applications:

Samples Generated by Dropping Observations with Item Nonresponses: Often users confine their analysis to subsamples of respondents who provided valid answers to certain questions. In this case, a weighted mean will not represent the entire population, but rather those persons in the population who would have given a valid response to the specified questions. Item nonresponse due to refusals, don't knows, or invalid skips is usually quite small, so the degree to which the weights are incorrect is probably quite small. In the event that item nonresponse constitutes a small proportion of the variables under analysis, population estimates (i.e., weighted sample means, medians, and proportions) would be reasonably accurate. However, population estimates based on data items that have relatively high nonresponse rates—such as family income—may not necessarily be representative of the underlying population of the cohort under analysis.

Data from Multiple Waves: Because the weights are specific to a single wave of the study, and because respondents occasionally miss an interview but are contacted in a subsequent wave, a problem similar to item nonresponse arises when the data are used longitudinally. In addition, occasionally the weights for a respondent in different years may be quite dissimilar, leaving the user uncertain as to which weight is appropriate. In principle, if a user wished to apply weights to multiple wave data, weights would have to be recomputed based upon the persons for whom complete data are available. In practice, if the sample is limited to respondents interviewed in a terminal or end point year, the weight for that year can be used.

Regression Analysis: A common question is whether one should use the provided weights to perform weighted least squares when doing regression analysis. Such a course of action may lead to incorrect estimates. If particular groups follow significantly different regression specifications, the preferred method of analysis is to estimate a separate regression for each group or to use indicator variables to specify group

membership; regression on a random sample of the population would be misspecified. Users uncertain about the appropriate method should consult an econometrician, statistician, or other person knowledgeable about the data before specifying the regression model.

User Notes: The NLSY97 data set contains two sampling weight variables for each survey round: SAMPLING_WEIGHT and CS_SAMPLING_WEIGHT. The first set includes all NLSY97 respondents. These weights (when divided by 100) will add up to an estimate of the number of U.S. residents in the sample age range in 1997. The second set contains weights only for respondents in the cross-sectional sample; all oversample cases have a zero weight. These weights are also designed to produce an estimate of the number of U.S. residents in the sample age range. Since there are fewer respondents if the oversample is omitted, however, each black or Hispanic respondent in the cross-sectional sample has a larger value.

For research that includes analysis by race, using the regular sampling weights rather than the cross-sectional weights will produce results with higher precision for black and Hispanic youths. For research that focuses only on non-black, non-Hispanic youths or that does not include any analysis by race/ethnicity, using the cross-sectional weights will save processing time.

Design Effects

Because the samples are multi-stage stratified random samples instead of simple random samples, respondents tend to be clustered in geographic areas (for more information on the sample design and screening process, see section 2.1). In general, these clusters tend to be alike in a variety of ways for a variety of reasons. For example, there may be cultural differences by locality or ecological differences in labor market conditions. Depending upon the degree of this homogeneity, the conventionally computed standard deviations for the variables, which assume a simple random sample, may be too small. However, by controlling the rate at which particular strata are sampled, multi-stage stratified random samples can improve upon simple random samples. The ratio of the correct standard error to the standard error computed under the assumption of a simple random sample is known as the design effect. The *NLSY97 Technical Sampling Report* provides design effects for the various strata.

As respondents in the cohort get older, mobility may mix the respondents more uniformly through the country, reducing the clustering of the sample as well as the design effects. Many of the persons who started out in the same PSU will have moved to different areas and may no longer be affected by similar unobservable labor market conditions. As this occurs, the error terms in a regression will more closely approximate the standard error computed for a completely random sample. However, some correlation due to respondents coming from the same household or neighborhood will, almost surely, remain.

By examining the geocode data for the NLSY97, it may be possible to control for some of the environmental factors generating design effects or, if desired, to compute design effects based upon county or metropolitan area clusters.

References

Moore, Whitney; Pedlow, Steven; Krishnamurty, Parvati; and Wolter, Kirk. *National Longitudinal Survey of Youth 1997 (NLSY97) Technical Sampling Report.* Chicago: NORC, 2000.